Calibration and validation of a genetic regulatory network model describing the production of the gap gene protein Hunchback in *Drosophila* early development

R. Dilão and D. Muraro

December 22, 2009

Nonlinear Dynamics Group, Instituto Superior Técnico,
Av. Rovisco Pais, 1049-001 Lisbon, Portugal.

rui@sd.ist.utl.pt, muraro@sd.ist.utl.pt

Abstract

We fit the parameters of a differential equations model describing the production of gap gene proteins Hunchback and Knirps along the antero-posterior axis of the embryo of *Drosophila*. As initial data for the differential equations model, we take the antero-posterior distribution of the proteins Bicoid, Hunchback and Tailless at the beginning of cleavage cycle 14. We calibrate and validate the model with experimental data using single- and multi-objective evolutionary optimization techniques. In the multi-objective optimization technique, we compute the associated Pareto fronts. We analyze the cross regulation mechanism between the gap-genes protein pair Hunchback-Knirps and we show that the posterior distribution of Hunchback follow the experimental data if Hunchback is negatively regulated by the Huckebein protein. This approach enables to predict the posterior localization on the embryo of the protein Huckebein, and we validate with the experimental data the genetic regulatory network responsible for the antero-posterior distribution of the gap gene protein Hunchback. We discuss the importance of Pareto multi-objective optimization techniques in the calibration and validation of biological models.

KEYWORDS: Genetic regulatory networks, Hunchback-Knirps cross regulation, Huckebein.

1

1   Introduction

In the *Drosophila* egg, maternal mRNAs are placed near the poles of the oocyte by the mother's ovary cells, defining the antero-posterior axis of the embryo. Fertilization triggers the translation of these maternal mRNAs to proteins that regulate the expression of zygotic genes. Each of the zygotic genes is transcribed in certain regions of the embryo syncitial blastoderm, and the produced proteins act as transcription factors that regulate the expression of other zygotic genes.

After fertilization, the first 13 nuclear divisions occur without the organization of cellular membranes, giving rise to a syncitial blastoderm. The cytoplasmic membranes only become completely formed three hours after fertilization, in the interphase following the 14th mitotic cycle, just before the onset of gastrulation.

During the syncitial stage, the transcribed zygotic genes are divided in three main families: gap, pair-rule and segment polarity genes. The proteins resulting from their expression define broad segmentation patterns along the antero-posterior axis of the embryo. These segmentation patterns appear as protein gradients along the antero-posterior axis of the *Drosophila* embryo, [Frigerio *et al.*, 1986; Driever and Nüsslein-Volhard, 1988; Akam, 1987; Nüsslein-Volhard, 1992].

The proteins with origin in the maternal mRNAs form gradients along the antero-posterior axis of the embryo. In the beginning of cleavage cycle 14, proteins of maternal origin act as transcription factors for gap-genes, pair-rule and segment polarity genes.

There are several models aiming to describe proteins steady gradients in *Drosophila* early development. Some models are based on the hypothesis of protein diffusion along the antero-posterior axis of the embryo, [Houchmandzadeh *et al.*, 2005; Alves and Dilão, 2006], and other models are based on the diffusion of mRNA of maternal origin, [Dilão and Muraro, 2009; Dilão *et al.*, 2009]. The protein diffusion hypothesis is sometimes justified by the absence of cellular membranes during the first 14 cleavage cycles of the embryo, and has been proposed by Nusslein-Volhard and co-workers in the late eighties, [Driever and Nüsslein-Volhard, 1988]. The mRNA diffusion hypothesis is supported by the recent observation of the mRNA Bicoid gradient, [Spirov *et al.*, 2009], and the associated diffusion mechanism has been reported by [Cha *et al.*, 2001] that observed rapid saltatory movements in injected mRNA *bicoid* with dispersion but without localization. Another maternal mRNA (*nanos*) has shown diffusive like behavior, [Forrest and Gavis, 2003].

Here, we will be concerned with the calibration and validation of

the genetic regulatory network involving maternal proteins and the antero-posterior distribution of the gap genes Hunchback (HB) and Knirps (KNI) along the *Drosophila* embryo. One of the reasons for this study is that the regulation of the gradient of the HB protein in the posterior region of the embryo of *Drosophila* is poorly understood, Margolis *et al.* [1995].

In order to calibrate the genetic regulatory network describing the production of the gap genes HB and KNI, we make some remarks on the biological assumption of our approach.

1) Models assuming that proteins of maternal origin diffuse along the embryo need the additional assumption that these proteins are continuously produced and degraded. However this is unrealistic because: (i) Degradation has never been observed, [Kerszberg and Wolpert, 2007]; (ii) There are no proteins in the space around nuclei suggesting that protein do not diffuse, [Dilão and Muraro, 2009]; (iii) Protein diffusion models need a condition on continuous production of mRNA of maternal origin, [Houchmandzadeh *et al.*, 2005; Alves and Dilão, 2006], a feature that has never been observed. On the other hand, models based on mRNA diffusion do not show these unrealistic features, and are able to produce accurately gradients of proteins of maternal origin, [Dilão and Muraro, 2009; Dilão *et al.*, 2009]. Anyway, the steady states obtained with the protein and the mRNA diffusion models have the same functional form, (compare Alves and Dilão [2006] with Dilão and Muraro [2009]), implying that the methodology followed here is not sensitive to the assumption about the diffusion model for proteins or mRNA of maternal origin.

2) Threshold effects are important phenomena for the establishment of positional information in the embryo, Wolpert [1969]. It has been shown that the mass action law models and the associated conservation laws lead naturally to stable gradients along the embryo of *Drosophila*, without the need of *ad-hoc* threshold or diffusive effects at the level of gap gene expression. As a consequence, production models for gap gene proteins are simply described by ordinary differential equations models derived from the mass action law. Positional information is an emergent property of the mass action associated conservation laws. These biological assumptions have been tested qualitatively in Alves and Dilão [2005], Alves and Dilão [2006] and Dilão and Muraro [2009b].

3) The mechanism of protein production is described in two steps. In the first step, we describe the establishment of the steady gradients of proteins produced from mRNAs with maternal origin.

In the second step, we consider that the maternal proteins are transcription factors for the gap-gene proteins. In order to simplify the model equations and the number of parameters for the description of the gap gene protein production, we assume that maternal origin proteins are not consumed in the activation or repression of the gap gene proteins. In the case of the Hunchback protein, in a first step, we consider that the protein is produced from mRNA with maternal origin. In a second step, it is assumed that HB is zygotically produced. In the initial gap gene phase, the gap gene proteins other than Hunchback are assumed to have zero initial concentration.

This paper is organized as follows. In section 2.1, we fit the experimental data of maternal proteins Bicoid (BCD), Hunchback (HB) and Tailless (TLL) with the equations for the steady state of a reaction-diffusion based model. The biological assumptions made are the ones described above in 1). The experimental data were taken from the FlyEx database, [Poustelnikova *et al.*, 2004; Pisarev *et al.*, 2009], and the fits were obtained by an evolutionary search algorithm. In these fits, we reproduce accurately the experimental data for BCD, HB and TLL, and we determine along the antero-posterior axis of the embryo of *Drosophila* the initial localization of the mRNA of maternal origin.

In section 2.2, we introduce the graph of the genetic network associated with the production and the cross regulation of the gap gene proteins HB and Knirps (KNI) and we derive a mass action production model. Then, we describe the process of calibration of the parameters of the model with the experimental data. The technique for parameter estimation is based on genetic algorithms with single- and multi-objective search techniques. As one of the main goals of this paper is to analyze the cross regulation of zygotically produced HB and KNI proteins, we have two objectives to fulfill. In this context, we find a continuous set of parameter solutions or Pareto front of the two-objectives optimization problem. This Pareto front corresponds to all possible admissible solutions of the bi-objective optimization problem. From the biological point of view, all the parameter solutions on the Pareto front are admissible and they correspond to different instances of the model parameters. All these Pareto solutions are very close to the experimental data and this has been evaluated by the chi-squared tests.

In section 3, we describe the methodology of the multi-parameter fitting with evolutionary algorithms for one-objective and multi-objective optimization techniques. This section is essentially qualitative, describing the geometry and structure of the algorithms. All the computations are computationally involved and the programs are included in

the supplementary material to this paper, Dilão and Muraro [2009c]. Finally, in section 4, we discuss the main conclusions of the paper.

2 Results and Discussion

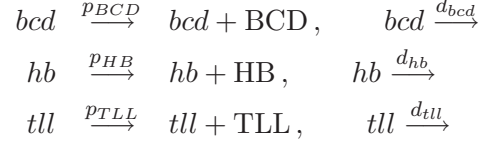## 2.1 Steady state models for the distribution of proteins with maternal origin

The first stages of the establishment of the positional information for the cellular differentiation of the *Drosophila* embryo are determined by the initial distribution of maternal mRNAs and the corresponding produced proteins. Here, we consider three proteins whose gradients are established prior to the gap gene phase. These three proteins are Bicoid (BCD), Hunchback (HB) and Tailless (TLL). We fit the steady state distribution of these proteins with the experimental data, taken from the FlyEx database [Kozlov *et al.*, 2000; Myasnikova *et al.*, 1999, 2001; Poustelnikova *et al.*, 2004; Pisarev *et al.*, 2009, http://flyex.ams.sunysb.edu/flyex/]. For the fits, we use a single-objective optimization technique for the distributions of BCD, HB and TLL.

Hunchback and *bicoid* maternal mRNA are initially distributed along the antero-posterior axis of the embryo. The *tailless* gene is activated by the Torso (TOR) protein that has maternal origin. Here, we consider that TLL is produced directly from mRNA *tll* which is not of maternal origin. This choice is a simplification in the model and the fit could be also obtained taking account of the activation of the *tll* gene by TOR, [Alves and Dilão, 2006].

To describe the steady states of BCD, HB and TLL, we assume a model for the production of proteins from the initial distribution of the associated mRNAs. In fact, we can adopt two alternative models. In one model, the produced protein diffuses and degrades along the embryo, leading to a gradient like steady state, [Alves and Dilão, 2006]. In a second alternative model, is the maternal mRNA that diffuses and degrades, leading to a gradient like steady state for the protein. The second model is experimentally supported by the fact the *bicoid* mRNA shows a gradient, [Spirov *et al.*, 2009]. It has been shown in Dilão and Muraro [2009] that the protein steady states for both models have the same functional form, with parameters assuming different biological meanings. In the following, and without lack of generality, we assume the simple mRNA diffusion model for the production of proteins of maternal origin (Assumption 1) in § 1.

In order to arrive at the steady state functional forms for the distribution of BCD, HB and TLL proteins along the antero-posterior

5

axis of the *Drosophila* embryo, we follow the mass action approach developed in Alves and Dilão [2005] and Dilão and Muraro [2009b]. We consider the following kinetic diagrams for protein production,

$$bcd \quad \overset{p_{BCD}}{\longrightarrow} \quad bcd + \mathrm{BCD}\,, \qquad bcd \overset{d_{bcd}}{\longrightarrow}$$
$$hb \quad \overset{p_{HB}}{\longrightarrow} \quad hb + \mathrm{HB}\,, \qquad hb \overset{d_{hb}}{\longrightarrow}$$
$$tll \quad \overset{p_{TLL}}{\longrightarrow} \quad tll + \mathrm{TLL}\,, \qquad tll \overset{d_{tll}}{\longrightarrow}$$

where capital letters represent proteins and the italic letters the corresponding mRNAs. The constants p_{BCD}, p_{HB} and p_{TLL} are the protein production rates from mRNAs, and d_{bcd}, d_{hb} and d_{tll} are mRNA degradation rates. By the mass action law, to the above kinetic diagrams correspond the equations for the concentrations,

$$\frac{\partial bcd}{\partial t} \;=\; -d_{bcd}bcd(x) + D_{bcd}\frac{\partial^2 bcd}{\partial x^2} \tag{1}$$

$$\frac{\partial \mathrm{BCD}}{\partial t} \;=\; p_{BCD}bcd(x) \tag{2}$$

$$\frac{\partial hb}{\partial t} \;=\; -d_{hb}hb + D_{hb}\frac{\partial^2 hb}{\partial x^2} \tag{3}$$

$$\frac{\partial \mathrm{HB}}{\partial t} \;=\; p_{HB}hb \tag{4}$$

$$\frac{\partial tll}{\partial t} \;=\; -d_{tll}tll + D_{tll}\frac{\partial^2 tll}{\partial x^2} \tag{5}$$

$$\frac{\partial \mathrm{TLL}}{\partial t} \;=\; p_{TLL}tll \tag{6}$$

This system of differential equations describe the production and distribution of proteins and mRNA along the antero-posterior axis of the embryo of *Drosophila*. The antero-posterior axis is described by the independent coordinate x. The x-dependent diffusion terms do not follow from the mass action law, but they have been added in order to describe the diffusive motion of the mRNAs. The diffusion constants of the mRNAs are D_{bcd}, D_{hb} and D_{tll}.

In order to solve this system of equations (1)-(6), we now define boundary and initial conditions. Denoting by L the length of the embryo, we have that $x \in [0, L]$. Assuming zero flux boundary conditions

for mRNAs and proteins, we have,

$$\frac{\partial bcd}{\partial x}(x=0,t)=0\,,\quad \frac{\partial bcd}{\partial x}(x=L,t)=0\,,\tag{7}$$

$$\frac{\partial \mathrm{BCD}}{\partial x}(x=0,t)=0\,,\quad \frac{\partial \mathrm{BCD}}{\partial x}(x=L,t)=0\tag{8}$$

$$\frac{\partial hb}{\partial x}(x=0,t)=0\,,\quad \frac{\partial hb}{\partial x}(x=L,t)=0\,,\tag{9}$$

$$\frac{\partial \mathrm{HB}}{\partial x}(x=0,t)=0\,,\quad \frac{\partial \mathrm{HB}}{\partial x}(x=L,t)=0\tag{10}$$

$$\frac{\partial tll}{\partial x}(x=0,t)=0\,,\quad \frac{\partial tll}{\partial x}(x=L,t)=0\,,\tag{11}$$

$$\frac{\partial \mathrm{TLL}}{\partial x}(x=0,t)=0\,,\quad \frac{\partial \mathrm{TLL}}{\partial x}(x=L,t)=0\tag{12}$$

for every $t \geq 0$. As initial conditions, we take the piecewise constant functions,

$$
\begin{aligned}
bcd(x,t=0) &= \begin{cases} B>0, & \text{if } 0<L_1<x<L_2<L \\ 0, & \text{otherwise} \end{cases}\\[2pt]
\mathrm{BCD}(x,t=0) &= 0\\[2pt]
hb(x,t=0) &= \begin{cases} H>0, & \text{if } 0<M_1<x<M_2<L \\ 0, & \text{otherwise} \end{cases}\\[2pt]
\mathrm{HB}(x,t=0) &= 0\\[2pt]
tll(x,t=0) &= \begin{cases} T_1>0, & \text{if } 0<N_1<x<N_2<N_3 \\ T_2>0, & \text{if } L_3<x<L_N<L \\ 0, & \text{otherwise} \end{cases}\\[2pt]
\mathrm{TLL}(x,t=0) &= 0
\end{aligned}
\tag{13}
$$

for every $x \in [0,L]$. The functions $bcd(x,t=0)$ and $hb(x,t=0)$ describe the distribution of bcd and hb maternal mRNA in the regions $[L_1, L_2]$ and $[M_1, M_2]$, respectively, of the antero-posterior axis of the embryo of *Drosophila*. The function $tll(x,t=0)$ is the distribution of the hypothetical *tll* maternal mRNA in the region $[N_1, N_2] \cup [N_3, N_4]$, and B, H, T_1 and T_2 are constants.

Equations (1)-(6), with boundary conditions (7)-(12), and initial conditions (13) define the mRNA diffusion model for BCD, HB and TLL production. This model is linear, and the steady states $\mathrm{BCD}_{eq}(x)$, $\mathrm{HB}_{eq}(x)$ and $\mathrm{TLL}_{eq}(x)$ can be obtained explicitly (for details see Dilão and Muraro

[2009]):

$$
\begin{aligned}
\text{BCD}_{eq}(x) &= 2\frac{a_1}{e^{2a_2/L} - 1}\cosh\left(a_2\frac{x}{L}\right)\left(\sinh\left(a_2\frac{L_2}{L}\right) - \sinh\left(a_2\frac{L_1}{L}\right)\right) \\
&+ \frac{a_1}{2}\left(e^{-a_2(x+L_1)/L} - e^{-a_2(x+L_2)/L}\right) + I_{bcd}(x) \qquad (14)
\end{aligned}
$$

$$
\begin{aligned}
\text{HB}_{eq}(x) &= 2\frac{a_3}{e^{2a_4/L} - 1}\cosh\left(a_4\frac{x}{L}\right)\left(\sinh\left(a_4\frac{M_2}{L}\right) - \sinh\left(a_4\frac{M_1}{L}\right)\right) \\
&+ \frac{a_3}{2}\left(e^{-a_4(x+M_1)/L} - e^{-a_4(x+M_2)/L}\right) + I_{hb}(x) \qquad (15)
\end{aligned}
$$

$$
\begin{aligned}
\text{TLL}_{eq}(x) &= 2\frac{a_5}{e^{2a_6/L} - 1}\cosh\left(a_6\frac{x}{L}\right)\left(\sinh\left(a_6\frac{N_2}{L}\right) - \sinh\left(a_6\frac{N_1}{L}\right)\right) \\
&+ \frac{a_5}{2}\left(e^{-a_6(x+N_1)/L} - e^{-a_6(x+N_2)/L}\right) + I_{1tll}(x) \\
&+ 2\frac{a_7}{e^{2a_8/L} - 1}\cosh\left(a_8\frac{x}{L}\right)\left(\sinh\left(a_8\frac{N_4}{L}\right) - \sinh\left(a_8\frac{N_3}{L}\right)\right) \\
&+ \frac{a_7}{2}\left(e^{-a_8(x+N_3)/L} - e^{-a_8(x+N_4)/L}\right) + I_{2tll}(x) \qquad (16)
\end{aligned}
$$

where,

$$
I_{bcd}(x) = \begin{cases}
\frac{a_1}{2}\left(e^{-a_2(L_1-x)/L} - e^{-a_2(L_2-x)/L}\right), & \text{if } x < L_1 \\
a_1 - \frac{a_1}{2}\left(e^{-a_2(x-L_1)/L} + e^{-a_2(L_2-x)/L}\right), & \text{if } L_1 \le x \le L_2 \\
\frac{a_1}{2}\left(e^{-a_2(x-L_2)/L} - e^{-a_2(x-L_1)/L}\right), & \text{if } x > L_2
\end{cases}
$$
(17)

$$
I_{hb}(x) = \begin{cases}
\frac{a_3}{2}\left(e^{-a_4(M_1-x)/L} - e^{-a_4(M_2-x)/L}\right), & \text{if } x < M_1 \\
a_3 - \frac{a_3}{2}\left(e^{-a_4(x-M_1)/L} + e^{-a_4(M_2-x)/L}\right), & \text{if } M_1 \le x \le M_2 \\
\frac{a_3}{2}\left(e^{-a_4(x-M_2)/L} - e^{-a_4(x-M_1)/L}\right), & \text{if } x > M_2
\end{cases}
$$
(18)

$$
I_{1tll}(x) = \begin{cases}
\frac{a_5}{2}\left(e^{-a_6(N_1-x)/L} - e^{-a_6(N_2-x)/L}\right), & \text{if } x < N_1 \\
a_5 - \frac{a_5}{2}\left(e^{-a_6(x-N_1)/L} + e^{-a_6(N_2-x)/L}\right), & \text{if } N_1 \le x \le N_2 \\
\frac{a_5}{2}\left(e^{-a_6(x-N_2)/L} - e^{-a_6(x-N_1)/L}\right), & \text{if } x > N_2
\end{cases}
$$
(19)

$$
I_{2tll}(x) = \begin{cases}
\frac{a_7}{2}\left(e^{-a_8(N_3-x)/L} - e^{-a_8(N_4-x)/L}\right), & \text{if } x < N_3 \\
a_7 - \frac{a_7}{2}\left(e^{-a_8(x-N_3)/L} + e^{-a_8(N_4-x)/L}\right), & \text{if } N_3 \le x \le N_4 \\
\frac{a_7}{2}\left(e^{-a_8(x-N_4)/L} - e^{-a_8(x-N_3)/L}\right), & \text{if } x > N_4
\end{cases}
$$
(20)

and,

$$a_1 = B\frac{p_{BCD}}{d_{bcd}}, \qquad a_2^2 = d_{bcd}\frac{L^2}{D_{bcd}} \tag{21}$$

$$a_3 = H\frac{p_{HB}}{d_{hb}}, \qquad a_4^2 = d_{hb}\frac{L^2}{D_{hb}} \tag{22}$$

$$a_5 = T_1\frac{p_{TLL}}{d_{tll}}, \qquad a_6^2 = d_{tll}\frac{L^2}{D_{tll}} \tag{23}$$

$$a_7 = T_2\frac{p_{TLL}}{d_{tll}}, \qquad a_8^2 = d_{tll}\frac{L^2}{D_{tll}} \tag{24}$$

Note that $a_6 = a_8$.

The steady states for the gradients of proteins BCD, HB and TLL are given by equations (14)-(24). For the calibration of equations (14)-(24) with the experimental data, we have taken from the FlyEx database the mean antero-posterior distributions of the proteins BCD, HB and TLL. These distributions have been calculated from the individual spatial distributions measured in 954 different embryos. These distributions are assumed to correspond to a steady state and, in the case of HB, the steady state is assumed to be established at the end of cleavage cycle 13. For the BCD and the TLL proteins, the steady state distribution corresponds to the beginning of cleavage cycle 14A. In Figs. 1, 2 and 3, we show the mean values and the corresponding standard deviations of the gradients of proteins BCD, HB and TLL along the antero-posterior axis of the embryo of *Drosophila*. In these pictures, all the embryos have been scaled to the length $L = 100$.

To fit the experimental data of BCD, HB and TLL with (14)-(24), we have used an evolutionary search algorithm (see §3.1), and the choice of parameters has been done by minimizing the reduced chi-square functions,

$$
\begin{aligned}
\chi^2_{BCD}(\vec{p_1}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{(\text{BCD}_{eq}(x_i,\vec{p_1})-\text{BCD}_{mean}(x_i))^2}{\text{BCD}_{\sigma^2}(x_i)} \\
\chi^2_{HB}(\vec{p_2}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{(\text{HB}_{eq}(x_i,\vec{p_2})-\text{HB}_{mean}(x_i))^2}{\text{HB}_{\sigma^2}(x_i)} \\
\chi^2_{TLL}(\vec{p_3}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{(\text{TLL}_{eq}(x_i,\vec{p_3})-\text{TLL}_{mean}(x_i))^2}{\text{TLL}_{\sigma^2}(x_i)}
\end{aligned} \tag{25}
$$

where $\vec{p_1} = (L_1, L_2, a_1, a_2)$ is the vector of the free parameters for the BCD production model, $\vec{p_2} = (M_1, M_2, a_3, a_4)$ is the vector of the free parameters for the HB production model, and

$$\vec{p_3} = (N_1, N_2, N_3, N_4, a_5, a_6, a_7, a_8)$$

9

is the vector of the free parameters for the TLL production model.
The functions $BCD_{mean}(x)$, $HB_{mean}(x)$ and $TLL_{mean}(x)$ are the mean
values of the protein concentrations along the antero-posterior axis
of the embryo, and the functions $BCD_{\sigma^2}(x)$, $HB_{\sigma^2}(x)$ and $TLL_{\sigma^2}(x)$
are the associated standard deviations. In the fits, we have assumed
that a_6 and a_8 are independent parameters and we have taken $n =$
100. This assumption gives more plasticity to the data fitting and
is based on the assumption that the goal of the fits is to find an
accurate fitting function for TLL. The protein TLL is activated by the
maternal origin protein Torso and this mechanism is not considered
here, [Alves and Dilão, 2006]. The results of the three calibrations are
shown in Figs. 1, 2 and 3, and the fitted parameter values are listed
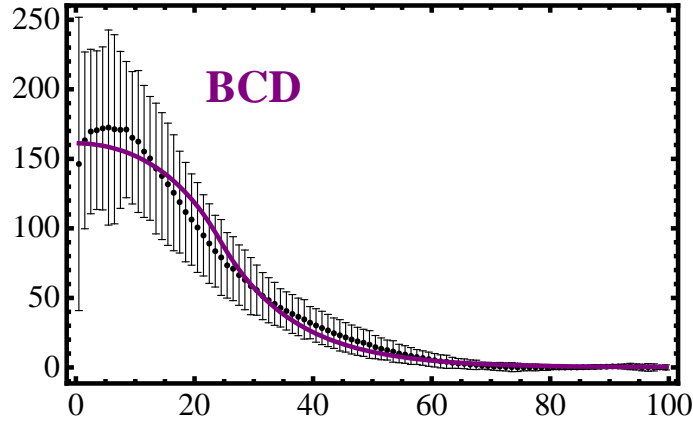in the figure captions.

Figure 1: Dots and error bars represent the mean values and the standard
deviations of the concentration of the protein Bicoid (BCD) along the antero-
posterior axis of the embryo of *Drosophila*, at cleavage cycle 14A. The fit has
been obtained with the steady state solution defined in (14), (17) and (21).
The parameter values found in the fit are: $L_1 = 0.00$, $L_2 = 0.24$, $a_1 = 186.83$
and $a_2 = 8.18$. The reduced chi-squared value of this fit is $\chi^2_{BCD}(\vec{p_1}) = 0.13$.
The interval $[L_1, L_2]$ is the region where mRNA *bcd* is deposited by the
mother's ovary cells.

From the fits in Figs. 1, 2 and 3, we conclude that the steady state
model describes well the distribution of proteins predicted from the
mRNAs with maternal origin. The values of the reduced chi-squared
test show that the agreement between data and fits are very good.
If a model is successfully calibrated with experimental data, then it
corresponds, with some degree of plausibility, to the mechanism that
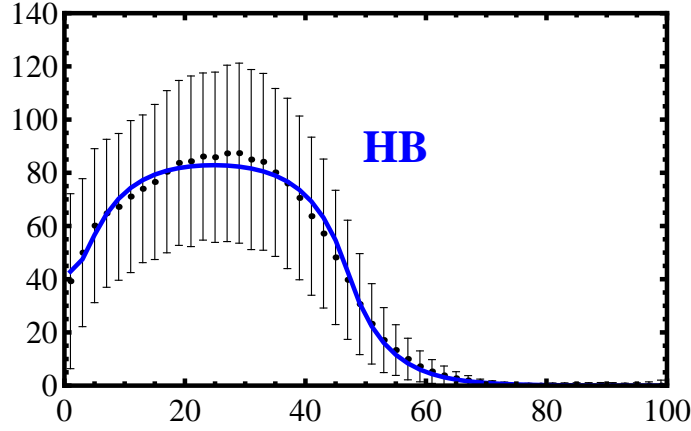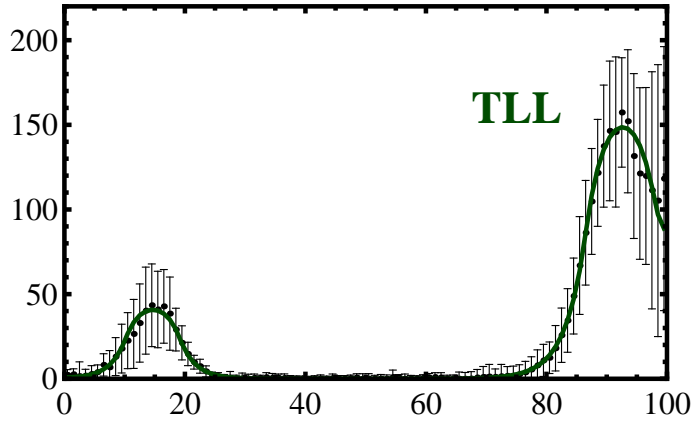it pretends to describe.

10

Figure 2: Dots and error bars represent the mean values and the standard deviations of the concentration of the protein Hunchback (HB) along the antero-posterior axis of the embryo of *Drosophila*, at the end of cleavage cycle 13. The fit has been obtained with the steady state solution defined in (15), (18) and (22). The parameter values found in the fit are: $M_1 = 0.04$, $M_2 = 0.47$, $a_3 = 85.09$ and $a_4 = 16.36$. The reduced chi-squared value of this fit is $\chi^2_{HB}(\vec{p_2}) = 0.02$. The interval $[M_1, M_2]$ is the region where mRNA *hb* is deposited by the mother's ovary cells.

Figure 3: Dots and error bars represent the mean values and the standard deviations of the concentration of the protein Tailless (TLL) along the antero-posterior axis of the embryo of *Drosophila*, at cleavage cycle 14A. The fit has been obtained with the steady state solution defined in (16), (19), (20), (23) and (24). The parameter values found in the fit are: $N_1 = 0.10$, $N_2 = 0.19$, $N_3 = 0.86$, $N_4 = 0.97$, $a_5 = 49.05$, $a_6 = 39.56$, $a_7 = 175.83$ and $a_8 = 30.33$. The reduced chi-squared value of this fit is $\chi^2_{TLL}(\vec{p_3}) = 0.03$.

11

As already stated, the steady state solutions (14)-(16) are functionally similar to the ones obtained with the protein reaction-diffusion model, compare equation (10) of Alves and Dilão [2006] with equation (4) of Dilão and Muraro [2009] .

Programs and software tools for evolutionary algorithms optimization techniques and model construction and analysis are available in the supplementary material, Dilão and Muraro [2009c].

We are now in condition to make the calibration and validation of the gap gene proteins HB and KNI.

2.2 Fitting the gap-genes

To describe the production of gap gene proteins, we consider that BCD, HB and TLL proteins are in the steady state with a gradient like distribution along the antero-posterior axis of the embryo of *Drosophila*, Figs. 1, 2 and 3. We consider that the production of the gap-genes proteins begins at the cleavage cycle 14 and, at this stage, we do not consider diffusion (Assumption 2) in § 1). We expect that the positional information is obtained by a threshold mechanism with diffusion playing no role, [Alves and Dilão, 2005; Dilão and Muraro, 2009b]. So, to model the gap-gene transcriptional regulation of Hunchback (HB) and Knirps (KNI), we take as initial conditions the antero-posterior distribution of BCD, HB and TLL, as found in the previous section. Then, we build the regulatory network following the mass action law strategy of Alves and Dilão [2005] and Dilão and Muraro [2009b].

The basic pattern of gap-genes HB and KNI expression pattern is due to strong mutual repression between these genes. This complementarity is particularly clear in the experimental data for the couple HB-KNI at cleavage cycle 14A-4, and has been confirmed in Jaeger and Reinitz [2006] and earlier results, together with the repression of TLL over KNI, affecting the posterior pole of the embryo.

The gap-gene genetic regulatory network involving HB and KNI is displayed in Fig. 4. Associated with the regulatory network of Fig. 4, we build the model for this genetic regulatory model based on the mass action law and following the description of transcriptional regulation by the operon model and developed in Alves and Dilão [2005] and Dilão and Muraro [2009b]. Using the *Mathematica* software package *GeneticNetworks.m*, we obtain the equations describing the time evolution of the gap gene protein concentrations. These differential equations involve the concentration of the proteins and of the gap genes with the different biding sites occupied or not. In the particular case of Fig. 4, the full system of ordinary differential equations has 14
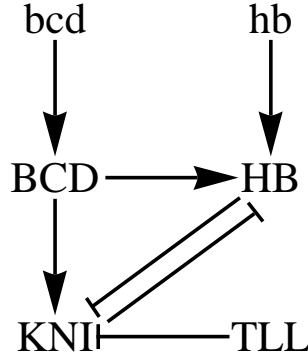
12

Figure 4: Genetic regulatory network graph associated with the cross regulation of the proteins Hunchback (HB) and Knirps (KNI) in *Drosophila* early development. The protein KNI is activated in the embryo by Bicoid (BCD). HB has a maternal origin and is also regulated by BCD. Both KNI and zygotically produced HB repress each other. In Fig. 2, we show the distribution of HB at the end of the maternal phase, before considering the regulation by BCD as described in this genetic network graph.

equations and 23 free parameters (Dilão and Muraro [2009c]).

In order to test the validity and completeness of the genetic regulatory network in Fig. 4, we took from the FlyEx database the experimental data of the distribution of HB and KNI for the late cleavage cycle 14, and we have integrated numerically in *Mathematica* the model equations generated by the *GeneticNetworks.m* software package. The free parameters on the model equations were determined with a bi-objective optimization technique (§3.3), minimizing the mean squared deviations between the model solutions and the experimental data. Denoting by $HB(x,t)$ and $KNI(x,t)$ the solutions of the model equations, we have fitted the experimental data for the antero-posterior distribution of HB and KNI with the functions $\alpha_{hb}HB(x,t)$ and $\alpha_{kni}KNI(x,t)$, where α_{hb} and α_{kni} are proportionality constants. The introduction of the proportionality constants α_{hb} and α_{kni} is due to the fact that experiments do not correspond to a direct measurement of local protein concentration, but it is proportional to protein concentration. These proportionality constants change from one protein to another. With these two additional proportionality constants and time as a free parameter, we have fitted the 23 parameters of the model with a bi-objective optimization technique and we have calculated the associated Pareto front.

In Fig. 5, we show this data and the corresponding fits. From the fitts, it is shown clearly that the genetic regulatory network of Fig. 4

13

describes well the HB and the KNI distributions away from the posterior tip of the *Drosophila* embryo. Clearly, complementarity of the proteins HB and KNI in the middle region of the embryo is observed. This fact suggests that there are other genes that regulate the posterior region of the embryo. A plausible candidate is the Huckebein (HKB) protein, Margolis *et al.* [1995].
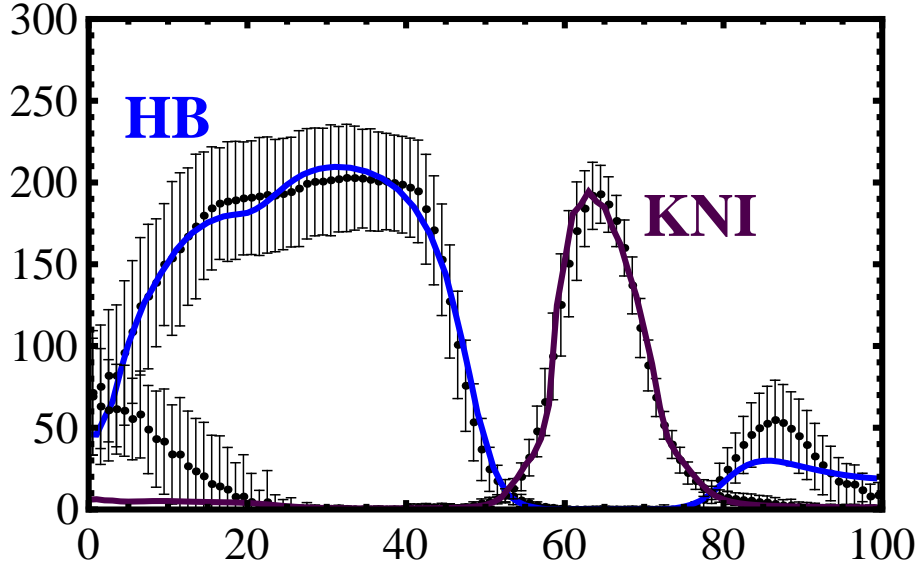

Figure 5: Dots and error bars represent the mean values and the standard deviations of the concentration of the proteins Hunchback (HB) and Knirps (KNI) along the antero-posterior axis of the embryo of *Drosophila*, at the end of cleavage cycle 14A. The fit has been obtained by a multi-objective optimization technique as described in § 3.3. The continuous lines correspond to the differential equation model solutions $\alpha_{hb}\mathrm{HB}(x,t^*)$ and $\alpha_{kni}\mathrm{KNI}(x,t^*)$, away from the steady state ($t^* < \infty$), and for a particular set of parameter values localized on the Pareto front of the bi-objective optimized solution. In this case, the fitted value of time is $t^* = 10$ s, and the fitted proportionality constants have the values $\alpha_{hb} = 0.1$ and $\alpha_{kni} = 2.0$. The penalized chi-squared values, (27), of these fits are $\chi^2_{HB}(\vec{p_4}) = 0.28$ and $\chi^2_{KNI}(\vec{p_4}) = 0.50$, where p_4 is the vector of the parameters that have been fitted. In this case, $P = 23$. In this model, this shows clearly that the genetic regulatory network of Fig. 4 describes well the HB and KNI distribution away from the posterior tip of the embryo of *Drosophila*.

In order to analyze the distribution of HB near the posterior region of the embryo, there is experimental evidence that Huckebein (HKB) protein has a band near the posterior pole of the embryo, repressing

14

the zygotic production of HB. Therefore, we introduce HKB in the gap gene regulatory network as in Fig. 6. As there is experimental evidence that HKB represses the production of HB near the posterior pole of the embryo, we assume a band type localization of HKB near the posterior tip of the embryo.
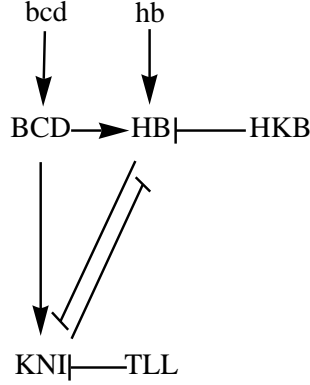

Figure 6: Genetic regulatory network graph associated with the cross regulation of the proteins Hunchback (HB), Knirps (KNI) and Huckebein (HKB) in *Drosophila* early development. The transcription repression of HKB on the transcription of HB is described in Margolis *et al.* [1995].

By consistence with the model construction done in the previous section § 2.1, we assume that the HKB protein is localized with the following steady state distribution,

$$
\begin{aligned}
\text{HKB}_{eq}(x) &= 2\frac{b_1}{e^{2b_2/L} - 1}\cosh\left(b_2\frac{x}{L}\right)\left(\sinh\left(b_2\frac{P_2}{L}\right) - \sinh\left(b_2\frac{P_1}{L}\right)\right) \\
&+ \frac{b_1}{2}\left(e^{-b_2(x+P_1)/L} - e^{-b_2(x+P_2)/L}\right) + I_{bcd}(x) \quad (26)
\end{aligned}
$$

where P_1, P_2, b_1 and b_2 are constants to be fitted and have the same meaning as the constants in the BCD equilibrium distribution (14). Under these conditions, we have introduced the HKB distribution into the previous model and we have repeated the bi-objective optimization analysis and we have calculated the associated Pareto front. In Fig. 7, we show one of the Pareto instances of the fit of the experimental data, as well as the fitted distribution of the protein HKB.

The quality of the fits in Figs. 5 and 7 were evaluated from the penalized chi-square functions,

$$
\begin{aligned}
\chi^2_{HB}(\vec{p}) &= \frac{1}{n - P/2}\sum_{i=1}^{n}\frac{(\text{HB}_{eq}(x_i, \vec{p}) - \text{HB}_{mean}(x_i))^2}{\text{HB}_{\sigma^2}(x_i)} \\
\chi^2_{KNI}(\vec{p}) &= \frac{1}{n - P/2}\sum_{i=1}^{n}\frac{(\text{KNI}_{eq}(x_i, \vec{p}) - \text{KNI}_{mean}(x_i))^2}{\text{KNI}_{\sigma^2}(x_i)}
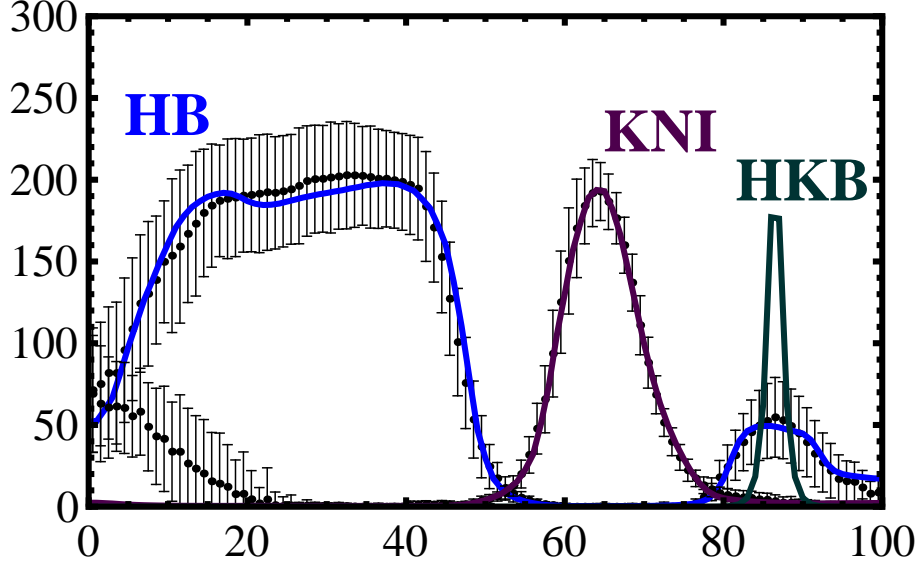\end{aligned} \quad (27)
$$

15

Figure 7: Dots and error bars represent the mean values and the standard deviations of the concentration of the proteins Hunchback (HB) and Knirps (KNI) along the antero-posterior axis of the embryo of *Drosophila*, at the end of cleavage cycle 14A. Due to the lack of experimental data on HKB its spatial experimental distribution is not represented. The fit has been obtained by a bi-objective optimization technique for HB and KNI, having also as free the parameters that describe the HKB distribution (26). The continuous lines correspond to the differential equation model solutions $\alpha_{hb}\text{HB}(x,t^*)$, $\alpha_{kni}\text{KNI}(x,t^*)$ and $\text{HKB}_{eq}(x)$, for a particular set of parameter values localized on the Pareto front of the bi-objective optimized solution. In the case, the fitted value of time is $t^* = 29.1$ s, and the fitted proportionality constants have the values $\alpha_{hb} = 0.11$ and $\alpha_{kni} = 0.65$. The penalized chi-squared value, (27), of this fit are $\chi^2_{HB}(\vec{p_5}) = 0.14$ and $\chi^2_{KNI}(\vec{p_5}) = 0.59$, where p_5 is the vector of the parameters that have been fitted. In this case, $P = 31$. The parameter values found for the prediction of the HKB distribution (26) are: $P_1 = 0.856$, $P_2 = 0.873$, $b_1 = 296.74$ and $b_2 = 121.87$. The HKB distribution found in the fit shows the existence of a stripe of the protein HKB near the posterior pole of the embryo as suggested experimentaly. With this fit, it is clearly shown that the genetic regulatory network of Fig. 6 describes well the HB and KNI distribution along all the antero-posterior axis of the embryo of *Drosophila*.

where $\vec{p}$ is the vector of the free parameters for the differential equation model and P is the dimension of the vector $\vec{p}$.

From the fits in Fig. 7, we conclude that the transcriptional cross repression of HB over KNI and the transcriptional repression of HKB over HB describe well the spatial distributions of HB and KNI proteins along the antero-posterior axis of the embryo of *Drosophila*. This result also predicts the distribution of the protein HKB.

Another important conclusion common to both fits is that gap gene protein expression is a dynamic process with a very fast expression time, of the order of 30 s (Fig. 7). This expression time is calculated relative to the beginning of cleave stage 14A.

Programs and software tools for multi-objective optimization techniques and Pareto front solutions are available in the supplementary material, [Dilão and Muraro, 2009c].

3 Materials and Methods

In this section, we briefly describe the algorithms that we have applied to calculate the parameters that best fit the experimental data to the model equations generated by the *Mathematica* software package *GeneticNetworks.m*. These algorithms are based on the *Covariance Matrix Adaptation Evolution Strategy* (CMA-ES) approach, an evolutionary algorithm for black-box continuous optimization, [Hansen and Ostermeier, 2001; Hansen, 2008]. The first algorithm is for single-objective optimization, used in §2.1, and will be referred by CMA-ES. The second algorithm is the multi-objective version of CMA-ES, used in §2.2, and uses several CMA-ES processes together with a global Pareto-dominance based selection, [Igel *et al.*, 2007]. In a maximization or a minimization problem, there is a fitness function relative to which an optimization is found. In multi-objective optimization problems, there are several fitness functions, and in general when we optimize in order to a fitness function, we are worsening in order to the other fitness function. Pareto optimization is a way of obtaining optimal solutions that are not dominated in a certain sense by other solutions.

3.1 Single-objective optimization: CMA-ES

CMA-ES is an evolutionary algorithm that uses a population of μ parents to generate λ offspring, and deterministically selects the best μ of those λ offspring for the next generation.

To have an idea of the parameter identification search problem, we take first a compact subset X of the parameter space S. The number of the parameter to be identified is the dimension of S. Set

an initial point $p_0 \in X \subset S$ and let $C = I_n$ be a covariance matrix, where I_n is the $n \times n$ identity matrix. Then, from the multivariate Gaussian distribution with covariance matrix C and mean value p_0, sample λ offsprings. For each offspring calculate a fitness function, in our case the, chi-squared distributions (25). From the best μ ($< \lambda$) offsprings, according to the fitness function, recalculate a new mean value p_0 and a new (unbiased estimator) covariance matrix C, and repeat the procedure. After several iterations, the best individual ever found is a candidate for the best choice of parameters. For details see Hansen and Ostermeier [2001] and Hansen [2008]. Maternal protein distributions in Figs. 1, 2 and 3 have been determined according to this technique.

3.2 Pareto optimization

Pareto optimization is concerned with the finding of the set of *optimal trade-offs* between conflicting objectives. Namely, Pareto solutions of a multi-objective problem are optimized solutions such that the value of one objective cannot be improved without degrading the value of at least another objective. Such best compromises are what is called the *Pareto set* of the multi-objective optimization problem. Pareto optimization is based on the notion of *dominance*. Consider a minimization problem with M real valued objective functions $f = (f_1, \ldots, f_M)$ defined on a subset $X \subset \mathbb{R}^n$. A solution of the optimization problem $\bar{x} \in X$ is said to *dominate* another solution $x \in X$, denoted by $\bar{x} \prec x$, if,

$$
\begin{aligned}
\forall m &\in \{1, \ldots, M\} : \\
(f_m(\bar{x}) &\leq f_m(x)) \wedge (\exists m \in \{1, \ldots, M\} : f_m(\bar{x}) < f_m(x)) .
\end{aligned}
$$

The Pareto set of an optimization problem is the set of nondominated solutions of a minimization (maximization) problem. More formally,

$$
\text{Pareto Set} = \{x : (x \in X) \wedge \nexists \bar{x} \in X : \bar{x} \prec x\} .
$$

The Pareto front is the image of the Pareto set in the fitness space,

$$
\text{Pareto front} = \{f(x) : (x \in X) \wedge \nexists \bar{x} \in X : \bar{x} \prec x\} .
$$

The goal of Pareto optimization is to find the *Pareto set* of optimized parameters and the Pareto front. Therefore, in a multi-objective approach, the natural choice for unbiased parameter estimation is the determination of the Pareto set of a given optimization problem. In this set, all the solutions are optimized solutions. The distributions of the gap gene proteins HB and KNI in Figs. 5 and 7 correspond

to parameter values on a Pareto set of the bi-objective optimization problem. In general, all the solutions on the Pareto set are equally acceptable, [Dilão *et al.*, 2009].

3.3 Multi-objective optimization: MO-CMA-ES

The Multi-Objective CMA-ES (MO-CMA-ES) optimization technique is based on the specific CMA-ES algorithm with a random choice of a large number of initial points in the search parameter space, [Igel *et al.*, 2007] . Once defined the multidimensional parameter search space, X, we proceed with the multi-objective optimization technique to determine the Pareto set and Pareto front of the two fitting problems of § 2.2. The MO-CMA-ES techniques can be divided in three steps:

1) In the compact search space X, choose randomly μ parents. For each parent, one offspring is generated with the CMA-ES algorithm. Initially, the CMA-ES algorithm is implemented with the identity as covariance matrix.

2) We now rank the best μ individuals from the set of 2μ individuals found previously. For that we use the concept of Pareto dominance. From the 2μ individuals, we select the set of all the non-dominated individuals and we give them rank 1. We apply the same procedure to the remaining individuals and we obtain the rank 2 individuals, [Deb *et al.*, 2002]. This procedure continues until a last rank is reached.

3) In order to rank the individuals within the same rank of non-dominance found previously, we do a second ranking of individuals within each rank. This second order ranking is done according to an *hypervolume measure* in the objective space, [Knowles *et al.*, 2003]. After this new ranking, we retain only the best μ individuals. With this procedure, we obtain an approximation to the Pareto front with an approximately uniform distribution of individuals within each rank. Then, we repeat these three procedures until a good converge to the Pareto front is achieved.

In Fig. 8, we show the Pareto front for the bi-objective optimization problem associated with the parameter identification describing the distribution of HB and KNI as shown in Fig. 5. We show the position of the fit of Fig. 5 in the Pareto front of Fig. 8. In Fig. 9, we show two other instances of the fits of HB and KNI proteins on the Pareto front. Comparing the three fits, we conclude that they are all acceptable.

In Fig. 10, we show the Pareto front for the fit of Fig. 7 and we mark the particular instance of the parameters of Fig. 7. In all the cases shown here, we conclude that the experimental data are optimally realized by an infinite set of parameters. This is particularly important in biology in the case of selection pressure affecting simultaneously several phenotypic characteristics of organisms.
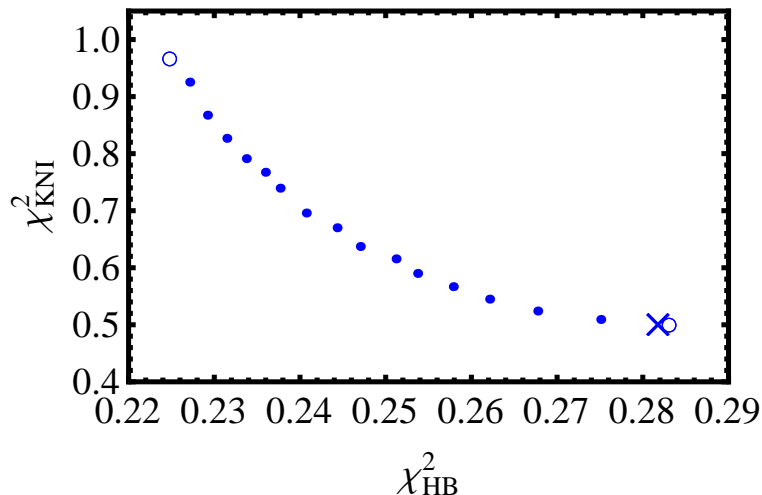

Figure 8: Pareto front for the fit of HB and KNI proteins of Fig. 5. In this bi-objective optimization problem, the coordinates of the fitness space are the reduced chi-squared functions $\chi^2_{HB}(\vec{p})$ and $\chi^2_{KNI}(\vec{p})$, where the vector of the parameters $\vec{p}$ is a parameterization of the Pareto front. These functions have been calculated as in (25). The cross represents the particular instance of the parameter values of Fig. 5. The circles represent the two other instances of the HB and KNI fits that are shown in Fig. 9.

4   Conclusions and Final Remarks

In order to analyze the expression of the gap gene protein Hunch-back along the antero-posterior axis of the embryo of *Drosophila*, we have introduced a genetic regulatory network model for the proteins HB and KNI and we have calibrated the experimental data with the model predictions. In the most complete version of the model of Fig. 6, we have shown that the distribution of HB and KNI along the antero-posterior axis of the embryo are in fact well described by a cross regulation mechanism together with the transcriptional repression of HKB over HB. We have predicted the distribution of HKB in the form of a localized stripe near the posterior tip of the embryo.
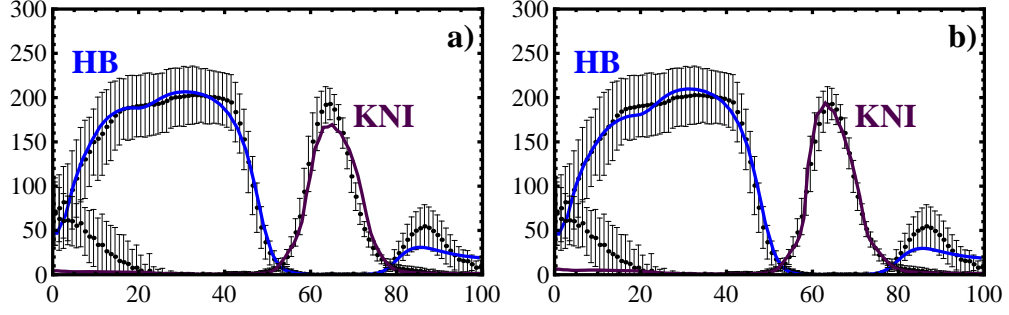
Figure 9: Two instances of the fit of HB and KNI in the Pareto front, represented by circles in Fig. 8. In a), we have the best fit for HB and the worst fit for KNI. IN b) we have the worst fit HB and the best fit for KNI.
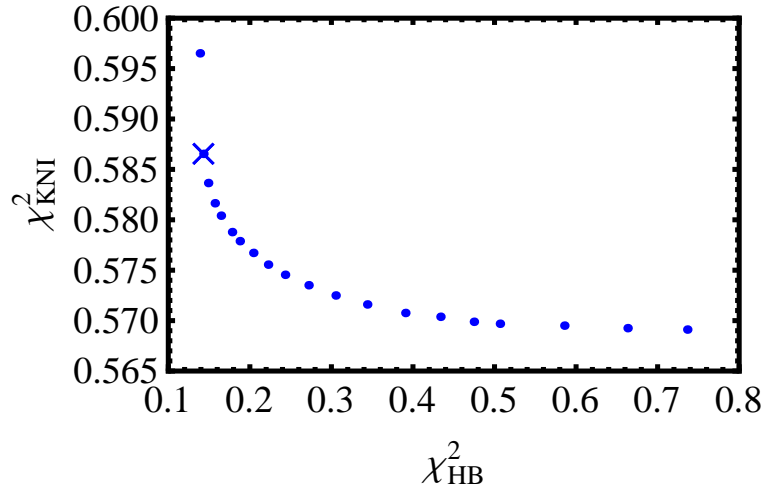

Figure 10: Pareto front for the fit of Hunchback, Knirps and Huckebein proteins of Fig. 7. The cross represents the particular instance of the parameter values of Fig. 7.

Another important conclusion we have obtained is that these patterns are obtained as transient solutions of an ordinary differential equation model, with diffusion playing no role at the level of gap gene protein expression patterns. With this approach, diffusion is only relevant for the establishment of gradients for proteins produced from mRNA with maternal origin. The patterning obtained along the embryo results from the differences in concentrations of the maternal proteins of the embryo.

The calibration and validation of the genetic regulatory network models have been done with genetic algorithm techniques for parameter identification. We have used single-objective and multi-objective techniques within the genetic algorithms formalism, and we have analyzed the usefulness of the concept of Pareto optimization in biology. Due to similarities between the fits and the experimental data, it is plausible to think that, in the presence of several objectives, the number of possible parametric solutions of a given problem is not unique, producing an infinite set of parameter instantiations. In this framework, the Pareto set and the Pareto front are the correct approach to analyze these problems. In the case of selection pressure on organisms affecting simultaneously several phenotypic characteristics, the Pareto type solutions appear as the right quantitative approach to quantify phenotypic variability.

In the most difficult case of the multi-objective optimization problem analyzed here, we have fitted 31 parameters in a system of ordinary differential equations with 18 independent variables, and we have implemented these algorithms in a grid computing environment. In the supplementary material of this paper, we list all the algorithms and all the associated C files developed under this framework, Dilão and Muraro [2009c]. These techniques are general and can be used in other parameter identification problems.

Acknowledgments

References

Akam, M., 1987. The molecular basis for metameric pattern in the *Drosophila* embryo. *Development* **101**, 1-22.

Alves, F. and Dilão, R., 2005, A simple framework to describe the regulation of gene expression in prokaryotes. *Comptes Rendus Biologies*, **328** 429–444.

Alves, F. and Dilão, R., 2006, Modelling segmental patterning in *Drosophila*: Maternal and gap genes. *Journal of Theoretical Biology*, **241** 342–359.

Cha, B.-J., Koppetsch, B. S., Theurkauf, W. E., 2001, In Vivo Analysis of *Drosophila bicoid* mRNA Localization Reveals a Novel Microtubule-Dependent Axis Specification Pathway. *Cell* 106, 35–46.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T., 2002, A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, **6**(2), 182–197. (2002).

Dilão, R. and Muraro, D., 2009, mRNA diffusion explains protein gradients in *Drosophila* early development. arXiv:0909.4239v1. Preprint.

Dilão, R. and Muraro, D., 2009b, Emergent thresholds in genetic regulatory networks: Protein patterning in *Drosophila* morphogenesis. arXiv:0909.4248v1. Pre-print.

Dilão, R. and Muraro, D., 2009c, Supplementary material for parameter calibration of genetic regulatory models. https://sd.ist.utl.pt/Download/download.html.

Dilão, R., Muraro, D., Nicolau, M. and Schoenauer, M., 2009, Validation of a morphogenesis model of *Drosophila* early development by a multi-objective evolutionary optimization algorithm. Proceedings of the 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining in BioInformatics, EvoBIO 2009, Springer Verlag, Tubingen, Germany, April 15-17, 2009.

Driever, W. and Nüsslein-Volhard, C., 1988, A gradient of bicoid protein in *Drosophila* embryos. *Cell* **54**, 83–93.

Forrest, K. M. and Gavis, E. R., 2003, Live imaging of endogenous RNA reveals a diffusion and entrapment mechanism for *nanos* mRNA localization in *Drosophila*. *Curr. Biol.* **13**, 1159-1168.

Frigerio, G., Burri, M., Bopp, D., Baumgartner, S. and Noll, M., 1986, Structure of the segmentation gene paired and the Drosphila PRD gene set as part of a gene network. *Cell* **47**, 735–746.

Hansen, N. and Ostermeier A., 2001, Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation* **9**, No. 2, 159–195.

Hansen, N., 2008, The CMA Evolution Strategy: A Tutorial. http://www.lri.fr/∼hansen/cmaesintro.html.

Houchmandzadeh, B., Wieschaus, E., Leibler, S., 2005, Precise domain specification in the developing *Drosophila* embryo. *Phys. Rev. E* **72**, 061920.

Igel, C., Hansen, N. and Roth, S., 2007, Covariance Matrix Adaptation for Multi-objective Optimization. *Evolutionary Computation* **15**, No. 1, 1–28.

Jaeger, J. and Reinitz, J., 2006, On the dynamic nature of positional information. *BioEssays* **28**, 1102–1111.

Kerszberg, M. and Wolpert, L., 2007, Specifying Positional information in the embryo:Looking Beyond Morphogens. *Cell* **130**, 205–209.

Knowles, J. D., Corne, D. W. and Fleisher, M., 2003, Bounded Archiving using the Lebesgue Measure. *Proceedings of CEC'2003*,**4**, 2490–2497, IEEE Press.

Kozlov, K., Myasnikova, E., Samsonova, M., Reinitz, J., Kosman, D., 2000 Method for spatial registration of the expression patterns of *Drosophila* segmentation genes using wavelets. *Computational Technologies* **5**, 112–119.

Margolis, J. S., Borowsky, M. L., Steingrímsson, E., Shim, C. W., Lengyel, J. A. and Posakony, J. W., 1995, Posterior stripe expression of hunchbackis driven from two promoters by a common enhancer element. *Development* **121**, 3067–3077.

Myasnikova, E., Kosman, D., Reinitz, J. and Samsonova, M., 1999, Spatio-temporal registration of the expression patterns of *Drosophila* segmentation genes. In Lengauer, T. & Schneider, R., Bork, P., Brutlag, D., Glasgow, J., Mewes, H.-W. and Zimmer, R. (eds), *Seventh International Conference on Intelligent Systems for Molecular Biology*, pp. 195–201. Menlo Park: AAAI Press.

Myasnikova, E., Samsonova, A., Kozlov, K., Samsonova, M. and Reinitz, J., 2001, Registration of the expression patterns of *Drosophila* segmentation genes by two independent methods. *Bioinformatics* **17**(1), 3–12.

Nüsslein-Volhard, C., 1992, Gradients that organize embryo development. *Scientific American* **275**(2), 54–61.

Pisarev, A., Poustelnikova, E., Samsonova, M. and Reinitz, J., 2009, FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic Acids Research* **37**, D560-D566.

Poustelnikova, E., Pisarev, A., Blagov, M., Samsonova, M. & Reinitz, J., 2004, A database for management of gene expression data in situ. *Bioinformatics* **20**, 2212–2221.

Rivera-Pomar, R. and Jäckle, H., 1996, From gradients to stripes in *Drosophila* embryogenesis: Filling in the gaps. *Trends Genet.* **12**, 478–483.

Spirov, A., Fahmy, K., Schneider, M., Frei, E., Nooll, M. and Baumgartner, S., 2009, Formation of the *bicoid* morphogen gradient: an mRNA gradient dictates the protein gradient. *Development* **136**, 605–614.

Wolpert, L., 1969, Positional information and the spatial pattern of cellular differentiation. *J. Theor. Biol.* **25**, 1–47.